Optimization

partial derivatives:

$(0,0,\cdots,\underset{\uparrow}{1},0,\cdots 0)$ (i^{th})

Defn. $f: \mathbb{R}^d \to \mathbb{R}$, $\{e_i\}_{i=1}^d$ standard basis, if

$$\lim_{h\to 0} \frac{f(a+he_i)-f(a)}{h} \text{ exists.}$$

we say f has partial derivative w.r.t x_i at a

$$\frac{\partial f(a)}{\partial x_i} \text{ or } \frac{\partial}{\partial x_i} f(a)$$

Treat everything else as a constant and take the 1-d derivative

Eg. let $a \in \mathbb{R}^d$, $f(x) = a^T x$. $\quad \frac{\partial f(x)}{\partial x_i} = \frac{\partial}{\partial x_i}\left(\sum_{j=1}^d a_j x_j\right) = \sum_{j=1}^d \frac{\partial}{\partial x_i} a_j x_j = a_i$

Gradient Fn. $f: \mathbb{R}^d \to \mathbb{R}$. Assume that partial derivatives of f all exist in a neighborhood of x and are also continuous at x. Then f is differentiable and its grad satisfies:

Under some conditions that will hold across this course.

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_d} \end{bmatrix}.$$

Convention $\quad \frac{\partial f}{\partial x} = \nabla f(x)^T$

Eg. l

Eg.

Eg. let $a \in \mathbb{R}^d$ and $f(x) = a^T x$ $\nabla f(x) = \left[\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \cdots, \frac{\partial f(x)}{\partial x_d}\right]^T$

$$= [a_1, a_2, \cdots, a_d]^T = a$$

Eg. let $W \in \mathbb{R}^{d \times d}$ and $f(x) = x^T W x$

$$[x_1, \cdots, x_d] \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1d} \\ w_{21} & w_{22} & \cdots & w_{2d} \\ \vdots & & \ddots & \vdots \\ w_{d1} & \cdots & & w_{dd} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \in \mathbb{R}$$

$$\left[\sum_{j=1}^{d} x_j w_{1j}, \sum_{j=1}^{d} x_j w_{2j}, \cdots, \sum_{j=1}^{d} x_j w_{dj}\right] \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

$$= \sum_{i=1}^{d} \sum_{j=1}^{d} x_i w_{ij} x_j = \sum_{i=1}^{d} w_{ii}^2 \cdot x_i + \sum_{i=1}^{d} \sum_{j \neq i} x_i w_{ij} x_j$$

$$\frac{\partial f(x)}{\partial x_m} = 2 x_m w_{mm} + \frac{\partial}{\partial x_m}\left(\sum_{i=1}^{d} \sum_{j \neq i} x_i w_{ij} x_j\right)$$

$$= 2 x_m w_{mm} + \frac{\partial}{\partial x_m}\left(x_m \sum_{j \neq m} w_{mj} x_j + \sum_{i \neq m} x_i \sum_{j \neq i} w_{ij} x_j\right)$$

$$= 2 x_m w_{mm} + \sum_{j \neq m} w_{mj} x_j + \sum_{i \neq m} x_i w_{im}$$

$$= \sum_{j=1}^{d} w_{mj} x_j + \sum_{i=1}^{d} x_i w_{im}$$

$$= (W[m,:])x + (W[:,m])x$$

$$= (W[m,:])x + (W^T[m,:])x$$

$$\begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_d} \end{bmatrix} = \begin{bmatrix} -(W[1,:]+\tilde{W}[1,:])^{\top} x \\ (W[2,:]+\tilde{W}[2,:])x \\ \vdots \\ (W[d,:]+\tilde{W}[d,:])x \end{bmatrix} = (W+\tilde{W})x.$$

Jacobian

Optimization Algos

1 Line search

for each step t

find a direction $u^{(t)}$

find $\alpha_t \in \mathbb{R}_+$ to minimize $h(\alpha) = f(w^{(t)} + \alpha u^{(t)})$

$$w^{(t+1)} = w^{(t)} - \alpha u^{(t)}$$

$u^{(t)}$ can be arbitrary coordinates $\Rightarrow$ coordinate descent.

If $u^{(t)}$ is the gradient direction $\Rightarrow$ gradient descent.

Newton's Method

$$w^{(t)} = \nabla f(w^{(t)})$$

Jacobian. A function $f \in \mathbb{R}^n \to \mathbb{R}^m$. $J_f(x) = \frac{\partial f}{\partial x} = \begin{pmatrix} \partial f_1/\partial x \\ \vdots \\ \partial f_m/\partial x \end{pmatrix} = (\partial f_i/\partial x_j)_{ij}$

$\frac{\partial f_i}{\partial x} = \frac{\partial (f(x))_i}{\partial x}$ cont. differentiable

Eg. $f: \mathbb{R}^d \to \mathbb{R}^k$ $f(x) = Ax$, $A \in \mathbb{R}^{k \times d}$. $Ax \in \mathbb{R}^k$

$$\frac{\partial f}{\partial x} = \begin{pmatrix} \partial (Ax)_1/\partial x \\ \partial (Ax)_2/\partial x \\ \partial (Ax)_k/\partial x \end{pmatrix} = \begin{pmatrix} \partial (A[1,:]x)/\partial x \\ \partial (A[2,:]x)/\partial x \\ \partial (A[k,:]x)/\partial x \end{pmatrix} = \begin{pmatrix} A[1,:] \\ A[2,:] \\ A[k,:] \end{pmatrix} = A$$

Chain rule. $f: \mathbb{R}^n \to \mathbb{R}^m$, $g: \mathbb{R}^m \to \mathbb{R}^p$. $g \circ f: \mathbb{R}^n \to \mathbb{R}^p$, f, g differentiable.

$J_{g \circ f}(x) = J_g(f(x)) J_f(x)$

Pf: $[J_{g \circ f}(x)]_{ik} = \frac{\partial (g \circ f)_i}{\partial x_k} = \sum_{j=1}^{m} \frac{\partial g_i(f(x))}{\partial f_j(x)} \cdot \frac{\partial f_j(x)}{\partial x_k} = \sum_{j=1}^{m} (J_g(f(x)))_{ij} [J_f(x)]_{jk}$

$= [J_g(f(x)) J_f(x)]_{ik}$

$= \frac{\partial g(f(x)_1, f(x)_2, \cdots, f(x)_m)}{\partial x_k}$

Hessian. twice differentiable $f: \mathbb{R}^d \to \mathbb{R}$ $\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{pmatrix}$

Usually, $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$

If 2nd order deriv. cont. then symmetric.

$f(x) = x^T A x$

$\nabla f(x) = (A + A^T) x = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_d} \end{pmatrix}$ $\frac{\partial \left(\frac{\partial f}{\partial x_i}\right)}{\partial x_j} = \frac{\partial^2 f}{\partial x_i \partial x_j} = A[j,:] + A^T[j,:]$ Taylor's theorem ⟹

$\nabla^2 f(x) = A + A^T$

Taylor's theorem (Multivariate case)

$f: D \to \mathbb{R}$ where $D \subseteq \mathbb{R}^n$ is a convex set

$$f(x) = f(u) + \nabla f(u)^T (x-u) + (x-u)^T \nabla_f^2(\xi)(x-u) + o(\|x-u\|^2)$$

Go back to Optimization Algs.

Newton's method:

$$\bar{f}(w) = f(w^{(t)}) + \nabla f(w^{(t)})^T (w - w^{(t)}) + \frac{1}{2}(w-w^{(t)})^T \nabla^2 f(w^{(t)})(w-w^{(t)})$$

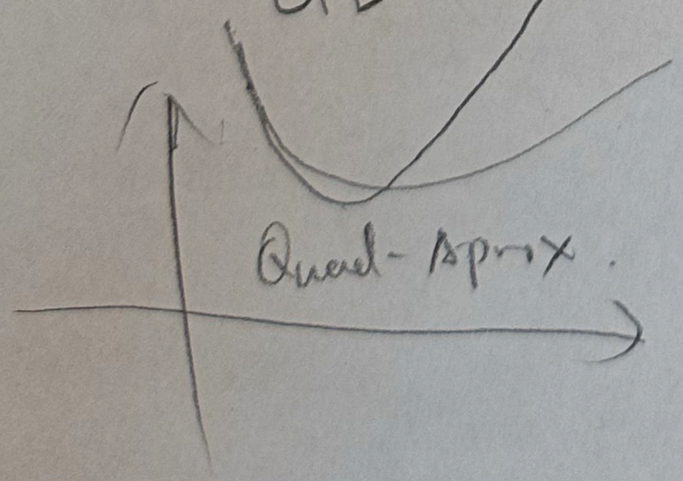
Quadratic function

then minimize $\bar{f}(w)$ $\Rightarrow$ take derivative, set to 0.
w.

$$\nabla \bar{f}(w) = \nabla f(w^{(t)})^T + 2 \cdot \frac{1}{2} \nabla^2 f(w^{(t)})(w - w^{(t)}) \overset{\text{set}}{=\!=\!=} 0$$

$$\text{Newton} \quad w^{(t+1)} = w^{(t)} - \left(\nabla^2 f(w^{(t)})\right)^{-1} \nabla f(w^{(t)})$$

$$\text{GD.} \quad w^{(t+1)} = w^{(t)} - \alpha_t \nabla f(w^{(t)}).$$



Another view of Newton's method 1-d case.

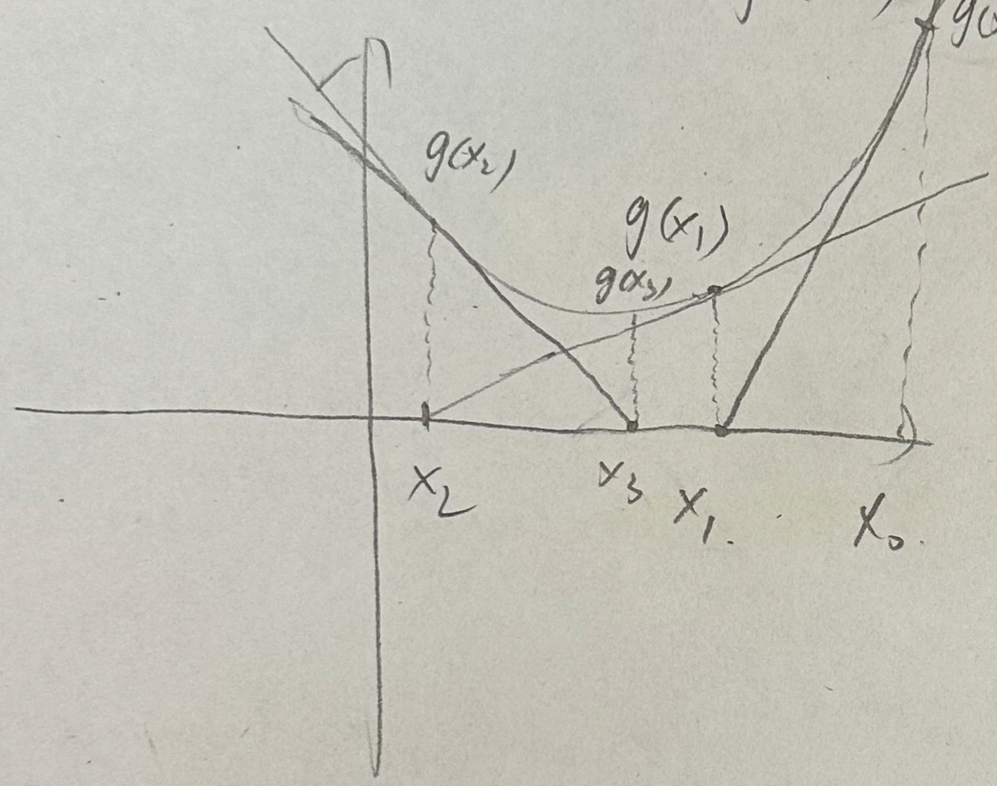
Want to find the critical pt., that is, find $\hat{w}$ s.t.

$$f(\hat{w}) = 0$$

let $g(w) = f'(w)$ linear approximation

$g(w) \approx g(w_0) + f'(w_0)(w - w_0) \Rightarrow$ find the zero-point of this lin-aprox.

$$w_1 = w_0 - \frac{f(w_0)}{f'(w_0)} = w_0 - \frac{g(w_0)}{g'(w_0)}$$



Eg $l_n(\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n} y_i \log(\mu(x_i^T\theta)) + (1-y_i)\log(1-\mu(x_i^T\theta))$

$$\nabla l_n(\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n} y_i \frac{1}{\mu(x_i^T\theta)} \cdot \mu(x_i^T\theta)(1-\mu(x_i^T\theta)) x_i + (1-y_i)\frac{1}{1-\mu(x_i^T\theta)} \cdot (-\mu(x_i^T\theta)(1-\mu(x_i^T\theta))) \cdot x_i$$

$$= -\frac{1}{n}\sum_{i=1}^{n} y_i (1-\mu(x_i^T\theta)) x_i + (1-y_i)(-\mu(x_i^T\theta)) x_i$$

$$= -\frac{1}{n}\sum_{i=1}^{n} (y_i - \mu(x_i^T\theta)) x_i$$

$$\nabla^2 \ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mu_i(x_i^T\theta)(1-x_i)x_i^T$$

$\mu_i(x_i^T\theta) > 0$ $\qquad$ $\nabla^2 \ell_n(\theta) \succeq 0$.

$\mu(1-\mu)$

$$\theta^{(t+1)} = \theta^{(t)} - \left(\frac{1}{n}\sum_{i=1}^{n} \mu_i(x_i^T\theta)x_i x_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}(\mu(x_i^T\theta) - y_i)\cdot x_i\right)$$

pros of Newton's method.

For certain fns. NR converges faster than GD way.

NR $|f(x^{(t)}) - f(x_*)| \leq \underbrace{\left(\frac{1}{2}\right)^{2^t - 1}}_{\leq \varepsilon}$

$t = O(\log(\log \frac{1}{\varepsilon}))$

GD $|f(x^{(t)}) - f(x_*)| \leq O(\frac{1}{t})$. $t \approx O(\frac{1}{\varepsilon})$.

Extreme example. If you have a Quadratic fn.

Newton's method converges in 1 step.

Cons of N-R.

3. sensitive to initial points. → need good initialization

4. calculating inverse takes $O(d^3)$ operations ⇒ single step slow.

2. unpredictable: even w/ very nice functions, eg. strongly convex & Lipschitz

it can still diverge.

Damped newton: $w_{t+1} = w_t - \eta_t (\nabla^2 f(w_t))^{-1} \nabla f(w_t)$

1 Hessian not invertible ⇒ regularization.